

Mathmatial Foundations of Reinforcement Learning

Online RL: regret analysis and algorithms



Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

Fall 2023

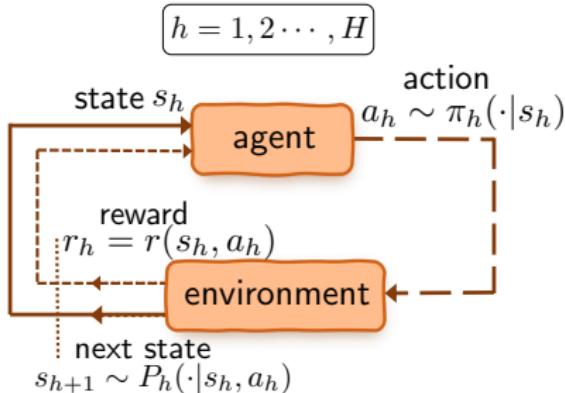
Outline

Episodic MDP and regret

Model-based RL with UCB exploration (UCB-VI)

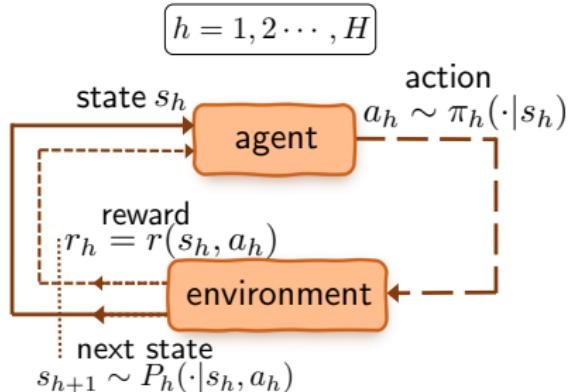
Model-free RL with UCB exploration (UCB-Q)

Finite-horizon nonstationary MDPs



- H : horizon length
- \mathcal{S} : state space with size S
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{h=1}^H$: policy (or action selection rule)
- $P_h(\cdot | s, a)$: transition probabilities in step h
- \mathcal{A} : action space with size A

Finite-horizon nonstationary MDPs

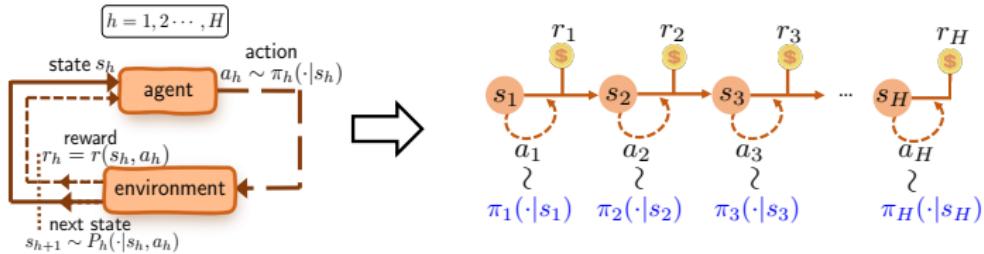


Value function: $V_h^\pi(s) := \mathbb{E} \left[\sum_{t=h}^H r_h(s_h, a_h) \mid s_h = s \right]$

Q-function: $Q_h^\pi(s, a) := \mathbb{E} \left[\sum_{t=h}^H r_h(s_h, a_h) \mid s_h = s, a_h = a \right]$



Bellman's optimality equation



Let $Q_h^*(s, a) = \max_\pi Q_h^\pi(s, a)$ and $V_h^*(s) = \max_\pi V_h^\pi(s)$.

- ① Begin with the terminal step $h = H + 1$:

$$V_{H+1}^* = 0, \quad Q_{H+1}^* = 0.$$

- ② Backtrack $h = H, H - 1, \dots, 1$:

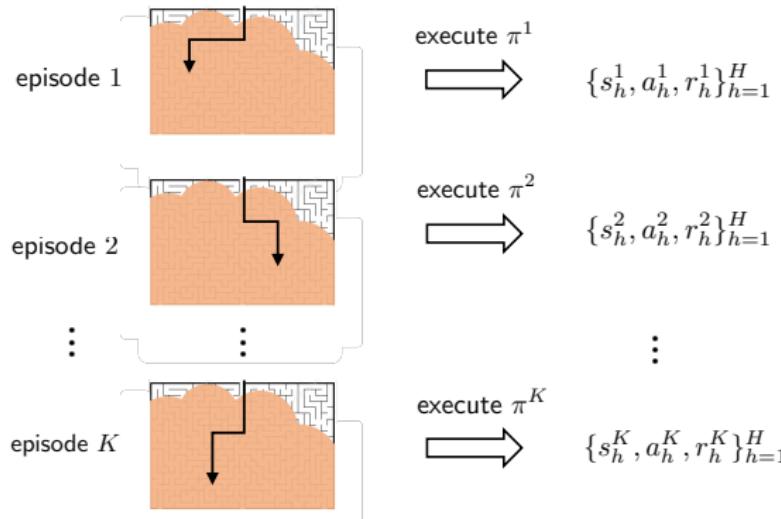
$$Q_h^*(s, a) := \underbrace{r_h(s_h, a_h)}_{\text{immediate reward}} + \underbrace{\mathbb{E}_{s' \sim P_h(\cdot | s, a)} V_{h+1}^*(s')}_{\text{next step's value}}$$

$$V_h^*(s) := \max_{a \in \mathcal{A}} Q_h^*(s, a), \quad \pi_h^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^*(s, a).$$

Online RL: interacting with real environments

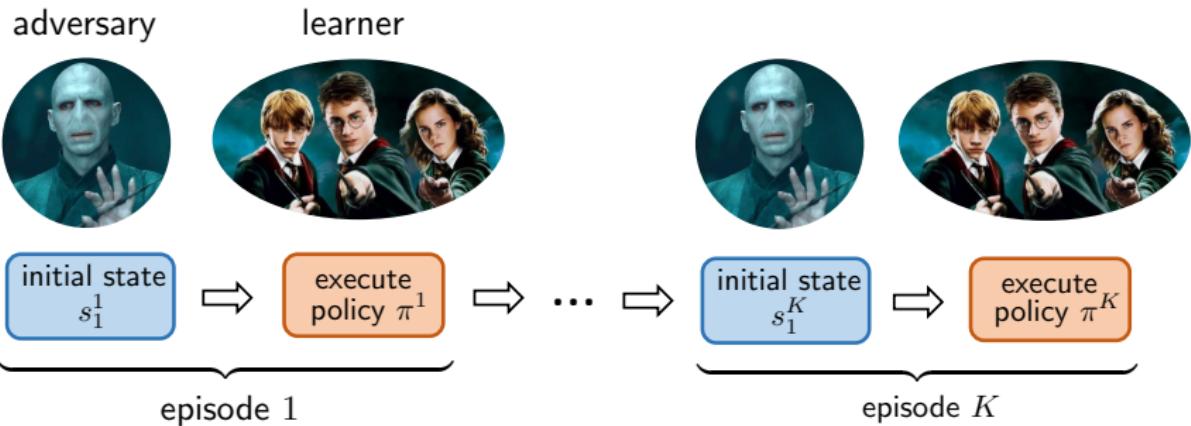
Sequentially execute MDP for K episodes, each consisting of H steps

— sample size: $T = KH$



exploration (exploring unknowns) vs. **exploitation** (exploiting learned info)

Regret: gap between learned policy & optimal policy



Performance metric: given $\underbrace{\{s_1^k\}_{k=1}^K}_{\text{chosen by nature/adversary}}$, define

$$\text{Regret}(T) := \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right)$$

Regret lower bounds

Theorem 1 ([Domingues et al., 2021])

For any algorithm, there exists an episodic MDP \mathcal{M}_π whose transitions depend on the stage h , such that for $T \geq H^2 SA$,

$$\mathbb{E}[\text{Regret}(T)] \geq \frac{1}{48\sqrt{6}} \sqrt{H^2 SAT}.$$

- Ignoring other factors, the regret is at least

$$\Omega(\sqrt{T}).$$

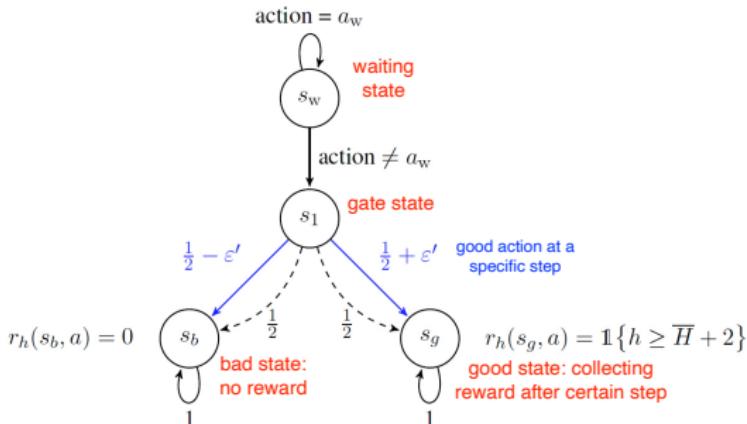
- The bound also reflects the impact of horizon length H and size of the state-action space SA . Note that the value function is on the order of H , so the “normalized” regret scales as

$$\frac{\mathbb{E}[\text{Regret}(T)]}{H} \gtrsim \sqrt{SAT} = \sqrt{SAHK}.$$

Construction of hard MDP

- Recall that the regret lower bound for an n -arm bandit (with normalized reward) is $\Omega(\sqrt{nT})$.
- It amounts to find a hard MDP that operates like a HSA -arm bandit (with reward $\sim H$).

Illustration of the hard MDP when $S = 4$. Taking $\bar{H} = \Theta(H)$.

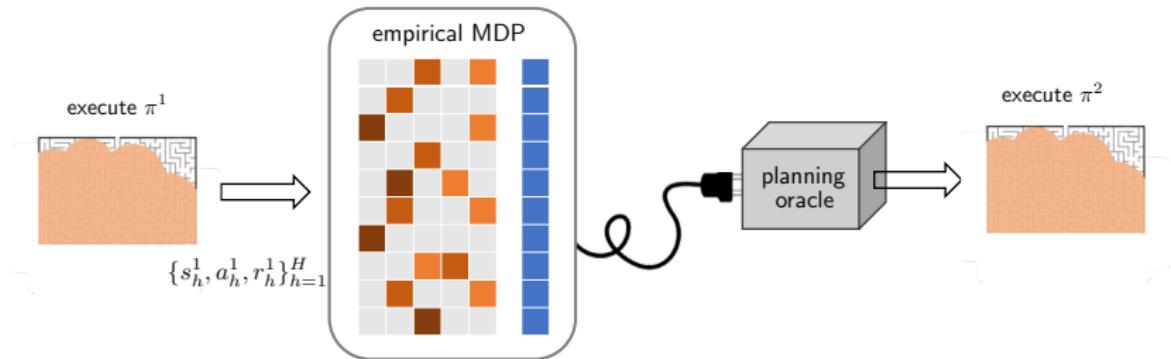


— Figure credit: [Domingues et al., 2021]

Can we design algorithms that achieve near-optimal regret?

Model-based RL with UCB exploration

Online RL with model-based approach



- Use **all** the previous data to estimate transitions (empirical frequencies)
- Apply planning (e.g., value iteration) on the estimated model to learn an updated policy for the next episode

How to balance exploration and exploitation in this framework?

UCB-VI: ideas

Motivated by the bandit UCB algorithm, [Azar et al., 2017] introduced **upper confidence bound (UCB)** into value iteration (VI).

- **Original VI:** Backtrack $h = H, H - 1, \dots, 1$:

$$Q_h(s, a) \leftarrow \underbrace{r_h(s_h, a_h)}_{\text{immediate reward}} + \underbrace{\hat{P}_{h,s,a} V_{h+1}}_{\text{next step's value}},$$
$$V_h(s) \leftarrow \max_{a \in \mathcal{A}} Q_h(s, a),$$

where $\mathbb{E}_{s' \sim P_h(\cdot | s, a)} V_{h+1}(s') = P_{h,s,a} V_{h+1}$ and $\hat{P}_{h,s,a}$ is the empirical estimate of $P_{h,s,a}$.

- Exploitation, but no exploration.
- Adding the UCB to $Q_h(s, a)$ similar to the bandit UCB algorithm.

UCB-VI: uncertainty quantification

Uncertainty in the next-step value $\hat{P}_{h,s,a} V_{h+1}$: recall that by Hoeffding's inequality and union bound, with probability at least $1 - \delta$,

$$\left\| (\hat{P}_{h,s,a} - P_{h,s,a}) V_{h+1}^* \right\|_\infty \lesssim \sqrt{\frac{H^2 \iota}{N_h(s, a)}},$$

where $N_h(s, a)$ is number of visits in (s, a) at step h .

Optimistic VI: run VI using rewards $\{r_h(s_h, a_h) + b_h(s_h, a_h)\}$

$$Q_h(s, a) \leftarrow \min \left\{ H - h + 1, \underbrace{r_h(s_h, a_h)}_{\text{immediate reward}} + \underbrace{\hat{P}_{h,s,a} V_{h+1}}_{\text{next step's value}} + \underbrace{b_h(s_h, a_h)}_{\text{bonus}} \right\},$$

$$V_h(s) \leftarrow \max_{a \in \mathcal{A}} Q_h(s, a),$$

where the bonus is $b_h(s_h, a_h) \asymp \sqrt{\frac{H^2 \iota}{N_h(s, a)}}.$

UCB-VI: algorithm

For each episode k :

- ➊ Backtrack $h = H, H - 1, \dots, 1$: run **optimistic value iteration**

$$Q_h(s, a) \leftarrow \min \left\{ H - h + 1, \underbrace{r_h(s_h, a_h)}_{\text{immediate reward}} + \underbrace{\hat{P}_{h,s,a} V_{h+1}}_{\text{next step's value}} + \underbrace{b_h(s_h, a_h)}_{\text{bonus}} \right\},$$

$$V_h(s) \leftarrow \max_{a \in \mathcal{A}} Q_h(s, a),$$

- ➋ Forward $h = 1, \dots, H$: take action according to the greedy policy

$$\pi_h(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(s, a)$$

and collect $\{s_h, a_h, r_h\}_{h=1}^H$.

Optimism in the face of uncertainty

Lemma 2 (Optimism)

With probability at least $1 - \delta$, it follows

$$Q_h^k(s, a) \geq Q_h^*(s, a), \quad V_h^k(s) \geq V_h^*(s)$$

for all (k, h, s, a) .

Optimism in the face of uncertainty:
acting according to $Q_h^k(s, a)$, which is an
upper bound of the true $Q_h^*(s, a)$.



Regret bound of UCB-VI with Hoeffding bonus

Theorem 3 ([Azar et al., 2017])

Let $\delta \in (0, 1)$. With probability at least $1 - \delta$, the regret of UCB-VI with Hoeffding bonus satisfies

$$\text{Regret}(T) \lesssim \sqrt{H^3 SAT\iota} + H^3 S^2 A \iota^3,$$

where $\iota = \log(HSAT/\delta)$.

- The regret bound scales as

$$\sqrt{H^3 SAT} \quad \text{as soon as} \quad T \gtrsim \underbrace{H^3 S^3 A}_{\text{burn-in cost}}.$$

which is sub-optimal by a factor of \sqrt{H} .

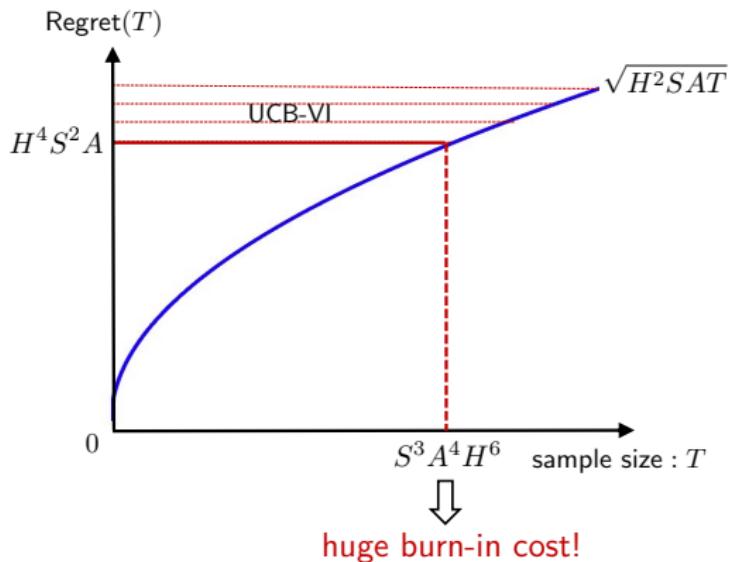
- By the optimism principle, the regret is bounded by

$$\text{Regret}(T) = \sum_{k=1}^K \left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \leq \sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \right).$$

Tighter UCB leads to smaller regret.

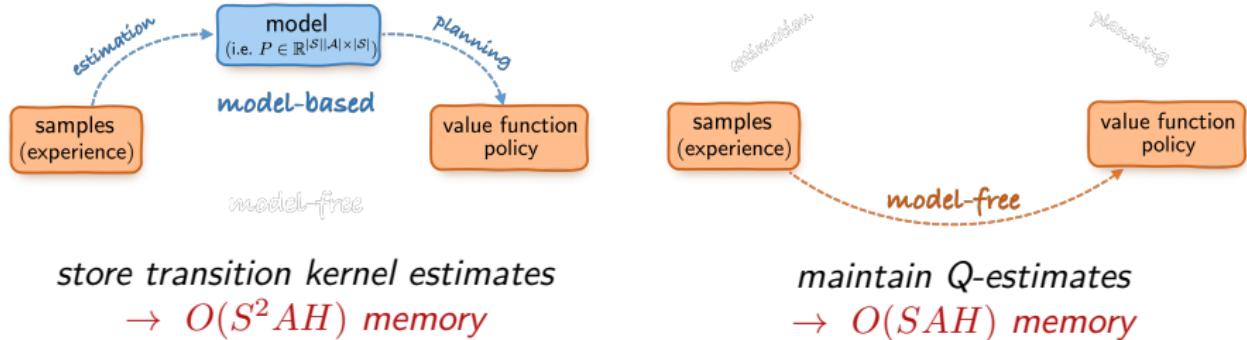
Near regret-optimal bound

By using tighter variance-aware concentration, [Azar et al., 2017] developed the first method that is asymptotically regret-optimal



Issues: (1) large burn-in cost; (2) large memory complexity
model-based: $S^2 A H$

Model-free RL is often more memory-efficient



Definition 4 ([Jin et al., 2018])

An RL algorithm is **model-free** if its space complexity is $O(SAH)$

Model-free RL with UCB exploration

Q-learning with UCB exploration

UCB-Q [Jin et al., 2018] modifies classical Q-learning with exploration bonus:
at the transition (s_h, a_h, s_{h+1})

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \alpha_t)Q_h(s_h, a_h) + \alpha_t (r(s_h, a_h) + V_{h+1}(s_{h+1}))}_{\text{classical Q-learning}} + \alpha_t \underbrace{b_h(s_h, a_h)}_{\text{bonus}}$$

- Using Hoeffding-type bonus to ensure the optimism property:

$$b_h(s, a) \asymp \sqrt{\frac{H^3 \iota}{N_h(s, a)}}$$

Large variability in stochastic update rules.

- Rescaled linear learning rates:

$$\alpha_t = \frac{H + 1}{H + t}, \quad t = N_h(s, a)$$

UCB-Q: algorithm with Hoeffding bonus

For each episode k :

- ① For $h = 1, \dots, H$:

- ① Take action according to the greedy policy $\pi_h(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q_h(s, a)$ and observe s_{h+1} ;
- ② Update the count $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$;
- ③ Compute the bonus $b_h(s_h, a_h)$;
- ④ Update the visited entry of Q -function:

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \alpha_t)Q_h(s_h, a_h) + \alpha_t (r(s_h, a_h) + V_{h+1}(s_{h+1}))}_{\text{classical Q-learning}} + \alpha_t \underbrace{b_h(s_h, a_h)}_{\text{bonus}}$$

- ⑤ Update value function:

$$V_h(s_h) \leftarrow \min\{H - h + 1, \max_a Q_h(s_h, a)\}.$$

Regret bound of UCB-Q with Hoeffding bonus

Theorem 5 ([Jin et al., 2018])

Let $\delta \in (0, 1)$. With probability at least $1 - \delta$, the regret of UCB-Q with Hoeffding bonus satisfies

$$\text{Regret}(T) \lesssim \sqrt{H^4 SAT\iota},$$

where $\iota = \log(HSAT/\delta)$.

- The regret bound

$$\sqrt{H^4 SAT}$$

is sub-optimal by a factor of H . No burn-in cost!

- Can be improved to $\sqrt{H^3 SAT}$ by using variance-aware concentration bounds (i.e., Bernstein inequality) to construct the UCB.

Can we design regret-optimal model-free algorithms?

Q-learning with UCB and variance reduction

[Zhang et al., 2020] incorporates **variance reduction** into UCB-Q:

$$\begin{aligned} Q_h(s_h, a_h) &\leftarrow (1 - \eta_k)Q_h(s_h, a_h) + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{UCB bonus}} \\ &+ \eta_k \left(\underbrace{\mathcal{T}_k(Q_{h+1}) - \mathcal{T}_k(\bar{Q}_{h+1})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q}_{h+1})}_{\text{reference}} \right)(s_h, a_h) \end{aligned}$$

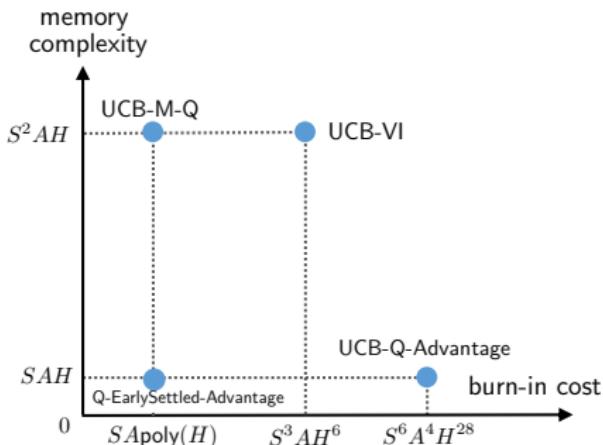
- Reference \bar{Q}_{h+1} , batch estimate $\hat{\mathcal{T}}$: help reduce variability

UCB-Q-Advantage is asymptotically regret-optimal

Issue: high burn-in cost $O(S^6 A^4 H^{28})$

Further developments on regret-optimal algorithms

Algorithm	Regret
UCB-VI [Azar et al., 2017]	$\sqrt{H^2SAT} + H^4S^2A$
UCB-Q-Advantage [Zhang et al., 2020]	$\sqrt{H^2SAT} + H^8S^2A^{\frac{3}{2}}T^{\frac{1}{4}}$
UCB-M-Q [Ménard et al., 2021]	$\sqrt{H^2SAT} + H^4SA$
Q-EarlySettled-Advantage [Li et al., 2021]	$\sqrt{H^2SAT} + H^6SA$



Model-free algorithms (Q-EarlySettled-Advantage) can simultaneously achieve
(1) regret optimality; (2) low burn-in cost; (3) memory efficiency

From regret to sample complexity

Question: given fixed initial state s_0 , how many samples does it take to find a policy $\hat{\pi}$ such that

$$V_1^*(s_0) - V_1^{\hat{\pi}}(s_0) \leq \varepsilon?$$

Note that the regret

$$\begin{aligned} \frac{1}{K} \text{Regret}(T) &= \frac{1}{K} \sum_{k=1}^K \left(V_1^*(s_0) - V_1^{\pi^k}(s_0) \right) \\ &= V_1^*(s_0) - \underbrace{\frac{1}{K} \sum_{k=1}^K V_1^{\pi^k}(s_0)}_{=: V_1^{\hat{\pi}}(s_0)}, \quad \text{where } \hat{\pi} \sim \text{Unif}(\{\pi_k\}_{k=1}^K). \end{aligned}$$

Setting $\frac{1}{K} \text{Regret}(T) \leq \varepsilon$ leads to $V_1^*(s_0) - V_1^{\hat{\pi}}(s_0) \leq \varepsilon$.

Example: regret of $\sqrt{H^2 SAT}$ leads to a sample size of $T = KH \gtrsim \frac{H^4 SA}{\varepsilon^2}$.

References I

-  Azar, M. G., Osband, I., and Munos, R. (2017).
Minimax regret bounds for reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. JMLR.org.
-  Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021).
Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited.
In *Algorithmic Learning Theory*, pages 578–598. PMLR.
-  Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018).
Is Q-learning provably efficient?
In *Advances in Neural Information Processing Systems*, pages 4863–4873.
-  Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021).
Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning.
Advances in Neural Information Processing Systems, 34.
-  Ménard, P., Domingues, O. D., Shang, X., and Valko, M. (2021).
UCB momentum Q-learning: Correcting the bias without forgetting.
In *International Conference on Machine Learning*, pages 7609–7618. PMLR.
-  Zhang, Z., Zhou, Y., and Ji, X. (2020).
Almost optimal model-free reinforcement learning via reference-advantage decomposition.
Advances in Neural Information Processing Systems, 33.